



Europäisches Patentamt
European Patent Office
Office européen des brevets



(11) Publication number : **0 519 670 A2**

(12)

EUROPEAN PATENT APPLICATION

(21) Application number : **92305481.1**

(51) Int. Cl.⁵ : **G11B 20/18, G06F 11/20**

(22) Date of filing : **15.06.92**

(30) Priority : **18.06.91 US 717263**

(43) Date of publication of application :
23.12.92 Bulletin 92/52

(84) Designated Contracting States :
DE FR GB IT

(71) Applicant : **International Business Machines Corporation**
Old Orchard Road
Armonk, N.Y. 10504 (US)

(72) Inventor : **Ng, Spencer Wah-Fung**
974 Hampswood Way
San Jose, California 95120 (US)
Inventor : **Palmer, David William**
248 Omira Drive
San Jose, California 95123 (US)
Inventor : **Thompson, Richard Scott**
67 Dearwell Way
San Jose, California 95138 (US)

(74) Representative : **Blakemore, Frederick Norman**
IBM United Kingdom Limited Intellectual Property Department Hursley Park
Winchester Hampshire SO21 2JN (GB)

(54) **Apparatus for maintaining fault tolerance in a disk drive array.**

(57) Fault tolerance in a redundant array 13 of disk drives 20, 21, 22, 23, 24 is degraded when errors exist in the array. Data in the array is rebuilt to remove the degradation, either for entire disk drives or for partial data, with minimal impact on array resources. Rebuilding occurs during idle time of the array or is interleaved between current data area accessing operations of the array at a rate which is inversely proportional to the activity level of the array. Alternatively, data is rebuilt when a data area being accessed is in need of rebuilding.

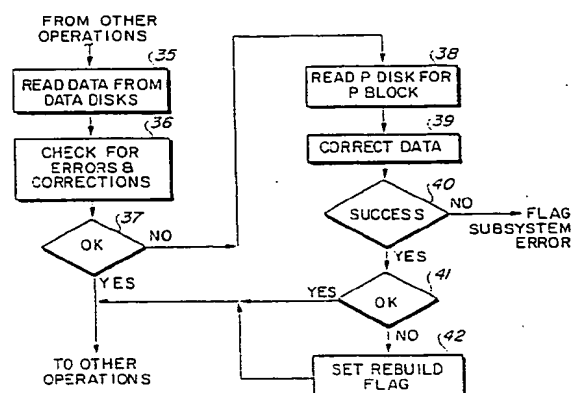


FIG. 3

EP 0 519 670 A2



The present invention relates to parity arrays of disk drives, and particularly to recovery from degraded redundancy in such an array by rebuilding data from error-affected tracks.

Patterson et al in the article "A CASE FOR REDUNDANT ARRAYS OF INEXPENSIVE DISKS (RAID)", ACM 1988, March 1988, describe several arrangements for using a plurality of data-storing disk drives. Various modes of operation are described; in one mode the data storage is divided among the several drives to effect a storage redundancy. Data to be stored is partially stored in a predetermined number of the disk drives in the array, at least one of the disk drives storing error detecting redundancies. For example, four of the disk drives may store data while a fifth disk drive may store parity based upon data stored in the four disk drives. Such a redundant array of disk drives may provide high data availability by introducing error correcting redundancy data in one of the disk drives. For example, four data blocks (one data block in each of the four drives) are used to compute an error correcting redundancy, such as a parity value; the computed error correcting redundancy is stored as a fifth block on the fifth drive. All blocks have the same number of data bytes and may be (not a requirement) stored in the five disk drives at the same relative track locations. The five drives form a parity group of drives. If any one of the drives in the parity group fails, in whole or in part, the data from the failing drive can be reconstructed using known error correcting techniques. It is desired to efficiently rebuild and replace the data from the failing disk drive while continuing accessing the drives in the array for data processing operations.

The disk drives in a parity group of drives may act in unison as a single logical disk drive. Such a logical drive has logical cylinders and tracks consisting of like-located cylinders and tracks in the parity group drives. In such array usage, the data being stored is partially stored in each of the data-storing drives in an interleaved manner in a so-called striped mode. Alternatively, the disk drives and their data in the parity group may be independently addressable and used in a so-called independent mode.

Whenever one of the disk drives in a single-parity array fails, even though data can be successfully recovered, the fault tolerance to error conditions is lost. To return to a desired fault tolerant state, the failing disk drive must be replaced or repaired and the affected data content rebuilt to the desired redundancy. It is desired to provide control means and methods for effecting such rebuilding of data and its redundancy to remove the error from a partially or wholly failed disk drive in a parity array of disk drives.

Accordingly the invention provides a method of automatically maintaining fault tolerance in a parity array consisting of a plurality of disk drives, comprising the machine-executed steps of:

detecting a degradation in the fault tolerance of the parity array;

evaluating the current information handling activity of the parity array;

establishing a plurality of data rebuild methods for removing the fault tolerance degradation from the parity array; and

analyzing the current information handling activity of the parity array and selecting one of the plurality of rebuild methods to ensure that the performance of said current information handling activity is not degraded below a predetermined level.

This approach allows a complete rebuild of data in a parity group of disk drives to a fault tolerant state after detecting loss or degradation of the fault tolerant state by a partially or wholly failed disk drive in a parity array of disk drives in a relatively non-intrusive manner while accesses to the parity array continue for data storage and retrieval.

Preferred examples of said plurality of data rebuild methods are:

(1) measuring the rate of machine operations of the array;

establishing the rate of rebuilding data affected by the error in inverse proportion to the measured rate of machine operations; and

rebuilding the data accordingly.

(2) detecting that the parity array is idle;

detecting that one of the disk drives is affected by error;

rebuilding the data during said detected idle times.

(3) performing a data area access operation to an addressable data unit in the parity array;

while performing the data area access operation, detecting that the addressed data unit is affected by error and in need of a data rebuild;

rebuilding the addressable data unit being accessed.

It is further preferred that the method includes:

completing a data rebuild using the rebuild method of (1) above;

detecting that the parity array is idle; and

continuing the data rebuilding of additional addressable error-affected data so long as the parity array is idle.

The invention also provides apparatus for automatically maintaining fault tolerance in a parity array consisting of a plurality of disk drives, the apparatus comprising:

means for detecting a degradation in fault tolerance in the parity array such that one of the disk drives needs to have data rebuilt;

means for measuring a rate of machine operations of said array;

means responsive to said measured operations rate for establishing a rate of rebuilding for the array for recovering from said degradation; and

a plurality of data rebuild means for performing data rebuild in said one disk drive.

It is preferred that the apparatus further comprises control means for controlling access to the disk drives in the parity array and for detecting when the array is currently not being accessed for a data handling operation. Preferred examples of data rebuild means are:

(1) data rebuild means connected to said control means and to said rate establishing means for determining whether a rebuild can be scheduled, and if so, activating the control means to give access to the parity array to perform a series of rebuild operations at the established rate.

(2) data rebuild means connected to said control means and responsive to the parity array not being currently accessed to activate the control means to give access to the parity array to perform a data rebuild.

(3) data rebuild means connected to the control means and responsive to a disk drive access occurring in an area of the array that needs a data rebuild to perform a data rebuild.

In example (1) above it is further preferred that the apparatus also includes means for indicating a plurality of rebuild rates, each corresponding to predetermined ranges of machine operations rates and in inverse proportion to the operations rate; and

means for indicating one of said plurality of rebuild rates as said established rate corresponding to the current machine operations rate.

Preferably each of the data rebuild means upon completing a data rebuild operation is responsive to the control means indicating no current access or no pending access requests to initiate another data rebuild operation.

Thus failures in a redundant array of disk drives are remedied by rebuilding the error-affected data using any one of a plurality of techniques to allow a continuing use of the disk drive array for information handling and data processing. One technique is a variable rate rebuild which schedules rebuilds at a rate in a detected inverse ratio to a current or pending rate of disk drive usage or accessing within a parity group. Upon completing each scheduled rebuild, this method and apparatus also preferably takes advantage of any idle time of the array by continuing rebuild if there is no waiting access. Another technique performs rebuild during predetermined array idle times by starting a non-scheduled rebuild of a predetermined portion of the error-affected data. A third or opportunistic technique detects a need for a data rebuild during a usual access to the array. The rebuilding may use any or all of the above techniques in conjunction with each other if desirable.

Thus data can be rebuilt onto a scratch or new disk drive which replaces a disk drive in error to restore redundancy of the array. The above techniques

also apply to a partially failed disk drive in which the error-affected data are rebuilt in a different track or zone of the disk drive in error; in the latter rebuild, data in the non-failing disk drives may also be moved to corresponding zones or tracks in the respective drives.

Thus one way of automatically maintaining fault tolerance of a redundant array of disk drives is by detecting that the fault tolerance of the parity array is degraded by a plurality of error-affected addressable data units of the parity array which respectively need data rebuilding to reestablish the fault tolerance, accessing the error affected disk drive, and during said access rebuilding data affected by said error. In other words, this involves performing a data area access operation in the parity array, and while performing the data area access operation, detecting that the data access operation is accessing one of the addressable data units needing a data rebuild and rebuilding at least the data unit being accessed.

Alternative approaches are: detecting that the array of disk drives has no current access and rebuilding data affected by said error affected disk drive beginning upon the detection of no current access; or determining the rate of information handling activity and selecting a data rebuild rate in a predetermined inverse ratio to the determined rate of information handling activity and using a variable rate rebuild which performs data rebuilding at said selected rebuild rate of a predetermined number of addressable error-affected data units in the parity array.

To recap therefore, a redundant disk array provides high data availability by introducing data redundancy such as parity check-sum across multiple drives. For example, four blocks of data of equal size from four different disk drives are used to compute a parity block of the same size and stored on a fifth drive. The five drives would form a parity group. With such an approach, when one drive in the array fails, its data can be reconstructed and made available from the other drives in its parity group. The drives in a parity group may act in unison as a single logical drive as a result of data being interleaved (striped) across the drives. This is called the striped mode. In addition, the drives in a parity group may be used independently. This is called the independent mode.

When one drive in a single-parity array fails, even though its data can be reconstructed through the other drives and is, therefore, logically available, the array is no longer in a fault tolerant state. A second drive failure within the parity group would render the data on the two failed drives lost or unavailable. To return to the fault tolerant state, the failed drive must be repaired and its content rebuilt.

Ideally, the data rebuild procedures should satisfy two objectives:

1. It should be completed as quickly as possible so as to minimise the window during which the ar-

ray is vulnerable to data loss/unavailability due to a second failure.

2. The rebuild procedure itself should be as non-intrusive as possible so as to minimise the impact on the performance of regular I/O services of the other disk drives in the array.

Because of objective 2, it is not acceptable to shut down the array and rebuild the data on the entire drive at one time, even though that would clearly satisfy objective 1. Rather, the content of the failed drive must be rebuilt piecewise. The present invention satisfies both objectives and it is applicable to arrays that operate in either the striped mode or the independent mode, or both.

When a unit of data is being rebuilt, a rebuild I/P must

1. read the corresponding data units from the other drives in the same parity group,
2. reconstruct the missing data unit by taking the exclusive-OR of the corresponding data units, then
3. write the reconstructed data unit to the drive being rebuilt.

A rebuild algorithm must answer the following three questions

- (a) how much data should be rebuilt at a time (i.e. for each rebuild I/O)?
- (b) when should a rebuild I/O be issued?
- (c) which part of the disk should the next rebuild I/O rebuild?

Rebuilding one block at a time appears to be inefficient, as the size of data rebuild unit (one block) is small relative to the time required to rebuild the unit. I/O performance, however, does not suffer, as the total duration of the rebuild I/O is short. In contrast, rebuilding one cylinder at a time may be very efficient, as the amount of rebuild work completed per rebuild I/O is large relative to the time required to rebuild the unit. Nevertheless, the cylinder rebuild ties up all the drives for a long period of time, severely degrading the performance of regular I/O's. Rebuilding one track at a time is a good compromise between the efficiency of the rebuild operation and its performance impact, although the invention will work with any reasonable size data rebuild unit.

Three ways are provided to generate a rebuild I/O.

Variable rate: First, rebuild I/O's can be generated at a variable rate depending on the current level of I/O activity. These are called scheduled rebuild I/O's. The array system monitors the rate of all regular I/O activities and adjusts the rebuild rate accordingly to maintain reasonable subsystem performance. A scheduled rebuild I/O does not have to wait for the array subsystem to be quiescent, but is injected into the subsystem periodically according to the rebuild rate prevalent at the time. This way, it can be ensured that rebuild activities do get scheduled even though the

subsystem may never be idle. A method for determining the rebuild I/O rate will be described later.

Furthermore, in order to make efficient use of the array subsystem's idle time, after a rebuild I/O has completed the rebuild of one track (or rebuild data unit) and there is no regular I/O waiting for any of the drives in the parity group, the rebuild I/O can be allowed to continue on to the next track until the next regular I/O arrives. When the next regular I/O arrives, the rebuild I/O can be immediately terminated (current track not rebuilt). Therefore, a scheduled rebuild I/O rebuilds a minimum of one track (or rebuild data unit) a time, but may rebuild optimally more if the subsystem has idle time.

When Idle: Second, when the array subsystem detects that there is no I/O activity on the affected parity group, it can generate a rebuild I/O. The rebuild I/O can be of one of two types, scheduled or non-scheduled. If the subsystem has not generated the required number of rebuild I/O's for the current time period, then it can generate the next scheduled rebuild I/O. This rebuild I/O works in exactly the same manner as described above, i.e. it is not interruptible until at least one track (or rebuild data unit) has been rebuilt. If the subsystem has already generated the required number of rebuild I/O's for the current time period, then it can generate a non-scheduled rebuild I/O. The non-scheduled rebuild I/O operates in a slightly different manner than the scheduled rebuild I/O. Since the required number of scheduled rebuild I/O's has been completed for the current time period, any extra non-scheduled rebuild I/O activity must not affect the performance of regular I/O activity. Therefore, the non-scheduled rebuild I/O is immediately terminated by the arrival of a regular I/O even though it may not have finished rebuilding the first track (or rebuild data unit) of the rebuild I/O. After completion of the rebuild of one track, the procedure can continue with the next track (or rebuild data unit) until terminated by the arrival of a regular I/O.

Opportunistic: In addition to scheduled and non-scheduled rebuilds, the array subsystem may generate rebuild I/O's when the opportunity presents itself. In this case, when a regular I/O requests data from the failed drive, the array subsystem rebuilds the entire track (or rebuild data unit) containing the requested data (if the requested data is larger than a track, then all of the tracks that contain the requested data are rebuilt). Since all the access mechanisms of the related drives are already in place to reconstruct the requested data, there is no additional overhead to reconstruct the entire track. Combining a regular I/O and a rebuild I/O into a single I/O has the effect of reducing the total I/O rate. Thus, the impact of rebuilding on the overall performance of the subsystem is minimised.

Again, in order to make efficient use of the array subsystem's idle time, at the end of a combined regular and rebuild I/O, if there is no regular I/O waiting

for any of the drives in the parity group, the rebuild I/O can be allowed to continue on to the next track until terminated by the arrival of the next regular I/O.

For an opportunistic rebuild I/O, where to rebuild is clear, viz., rebuild the track that the regular I/O is accessing. For a scheduled or non-scheduled rebuild I/O, it is preferred to select any track (or rebuild data unit) from the cylinder for which the maximum of the seeks of all the drives in the associated parity group is minimised. For a striped mode array, since the access mechanisms of all the drives in a parity group are always over one single logical cylinder, this would be the current-cylinder. For an independent mode array, let C_1, C_2, \dots, C_k be the current cylinders of the k drives in the parity group. Without loss of generality, if, $C_1 < C_2 < \dots < C_k$, then the rebuild cylinder C should be $(C_1 + C_k)/2$. If all the tracks in that cylinder have already been rebuilt, go to the adjacent cylinders ($C-1, C+1, C-2, \dots$ etc.).

To determine the rebuild I/O rate analytical modelling, trace-driven simulation, or actual measurement on a real system can be used to determine the overall performance (average response time vs. total I/O rate) of the array subsystem for each constant rebuild I/O rate.

One approach is to measure the average response time during each period of time directly. If the I/O request queues are maintained by the array subsystem, then this measurement is done at the subsystem level. If the I/O request queues are maintained on the host, then this measurement is done at the host level. In either case, the information is made known to the array subsystem which can then adjust the rebuild I/O rate accordingly.

In addition to rebuild I/O scheduling activity information, the array subsystem must keep track of which tracks have been rebuilt. This information is needed so that:

1. a rebuilt track will not be rebuilt a second time, and
2. if a regular I/O requests data from a track that has already been rebuilt, the data can be retrieved from that track directly without reconstruction from the other drives in the parity group.

This rebuild record keeping can be done in a straightforward manner using a bit map. This bit map will have one word for each cylinder of the disk drive; Word 1 corresponds to Cylinder 1, Word 2 corresponds to Cylinder 2, etc. Within a word, each bit will correspond to a track of that cylinder. Bit 1 corresponds to Track 1, Bit 2 corresponds to Track 2, etc. A "1" in a bit position indicates that track has not been rebuilt, while a "0" indicates that track has already been rebuilt. With such an arrangement, it can be quickly determined if a cylinder has already been completely rebuilt by testing the corresponding word for all 0's.

The invention therefore provides an efficient pro-

cedure for rebuilding the data on a failed disk in a redundant array, the salient features of which are:

1. The array system maintains an acceptable level of performance for regular I/O's during the rebuild period.
2. Rebuild activities will take place as long as an acceptable level of performance for regular I/O's can be maintained, thus guaranteeing that the rebuild will be completed within a reasonable amount of time.
3. Utilising the array's idle time whenever it becomes available, thus, reducing the rebuild's impact on the performance of regular I/O services, and shortening the total rebuild time.
4. Incorporating an intelligent way of choosing where to do the next rebuild so that it can be completed faster and more efficiently.
5. Combining rebuild activity with regular I/O service whenever possible, allowing some of the rebuild work to be done at little extra cost.

The invention will now be described in further detail by way of example with reference to the following drawings:

Fig. 1 illustrates in simplified form an information handling system employing the present invention.

Fig. 2 is a graph illustrating the principles of variable rate rebuilding in the Fig. 1 illustrated array of disk drives.

Fig. 3 is a machine operations chart showing detecting errors in the Fig. 1 illustrated array and priming the system for rebuilding.

Fig. 4 is a machine operations chart showing activating any one of three rebuild methods or apparatus.

Fig. 5 is a simplified machine operations chart showing selection of a rebuild method and apparatus.

Fig. 6 is a diagrammatic showing of a disk recording surface as may be used in practicing the present invention.

Fig. 7 is a diagrammatic showing of a data structure usable in practicing the present invention.

Fig. 8 is a diagrammatic showing of a bit map control for effecting rebuild.

Fig. 9 is a machine operations chart showing rebuild using a variable rate method and apparatus.

Fig. 10 is a machine operations chart showing rebuild using an array idle time method and apparatus.

Fig. 11 is a machine operations chart showing rebuild using the opportunistic rebuild method and apparatus.

Referring now more particularly to the appended drawing, like numerals indicate like parts and structural features in the various figures. Host processor(s) 10 (Fig. 1) are respectively connected to one or more controller(s) 11 by host to peripheral interconnection 12. A plurality of parity arrays 13, 14 and 15 are connected to controller 11 by a usual controller to peripheral device connection 17. Each of the Fig. 1 illustrat-

ed arrays 13-15 include five disk drives 20-24, no limitation thereto intended. Four of the disk drives 20-23 store like-sized blocks of data of one data unit. The block sizes may vary from data unit to data unit. A data unit can be an amalgamation of files, one file, graphic data, and the like. A fifth disk drive 24 is a parity or error detection redundancy storing drive P. The redundancy is a parity data block having the same size as the corresponding data blocks of the data unit. The redundancy is computed based upon any algorithm, including simple parity for example, using the data in the data blocks stored in drives 20-23. All data blocks of any data unit in disk drives 20-24 may be stored in the same relative tracks in all of the disk drives; i.e. all data blocks may be store in track 16 of all the drives, for example; while this storage arrangement simplifies data management, it is not essential as regards the present invention.

Disk drives 20-24 form a parity group, with disk 24 being dedicated to storing parity blocks. This general arrangement is known as RAID 3 and RAID 4 architecture, see Patterson et al, supra. Alternately, the storage of the parity block scan be rotated among all disk drives in the parity group with no single drive being designated as the parity drive. This latter storage arrangement is known as RAID 5 architecture. The present invention is equally applicable to any of these architectures, plus other architectures.

When storing data to any one or more of the disk drives 20-23, a new parity value is computed and stored in disk drive P 24. For efficiency purposes, it is desired to simultaneously record data in all four disk drives 20-23, and compute parity and record parity on disk drive P 24 as the data are being stored on the data drives; in a rotated parity arrangement, parity data are stored in the appropriate disk drive.

Host processors 10 and controller 11 both participate in the above-described machine operations; it is to be understood that the combination of host processors 10 and controller 11 comprises a computer means in which programming resides; such programming is represented by the machine operation charts of the various figures. The programming can be a separate part of the Fig. 1 illustrated system or can be embodied in ROM, loadable software modules, and the like.

Rebuilding data in a parity group on a disk drive that replaces a failed disk drive, either a spare or true replacement is achieved using a combination of approaches. Scheduled rebuilds are controlled by a variable rebuild rate method, an idle time rebuild occurs during any idle time of the parity array, and an opportunistic rebuild method is invoked upon each access to a replacement drive for the failed drive for accessing a non-built data storing area. This description assumes that a failed drive (eg a drive having a number of non-recordable data storing tracks/clusters of sectors, a failed mechanical part that prevents accessing

or the like) has been replaced with a scratch drive or disk using known disk drive and disk replacement procedures.

Before proceeding with a detailed description, the principle of the variable rate rebuild method is described with respect to Fig. 2. When this method is active, rebuild disk accesses (input/output operations or I/O's) are commanded or scheduled at a rate which varies inversely to the current level of I/O activity. Fig. 2 illustrates how the rate of rebuild scheduling is ascertained. Such rebuilding is interleaved with host processor 10 or other disk drive accesses, as will become apparent. A desired response time T is determined for the parity group to be managed. Such a response time is determined using known system analysis techniques, or the rate can be arbitrary and capricious. The five curves 30, also respectively labelled 1-5, show the variation of average response time (vertical ordinate) with the total I/O rate represented on the horizontal ordinate. The total I/O rate is determined by the activity of the host processors 10. The I/O rate is repeatedly monitored in predetermined constant measurement periods. The measured I/O rate determines the rebuild rate for the next ensuing measurement period. The measured rate during each measurement period is the computed average I/O rate for the measurement period of the parity group. When the I/O rate is higher than R1, then no rebuilds are scheduled during the next ensuing measurement period. During such a measurement period rebuilds may occur using either the idle or the opportunistic rebuild methods. The rebuild schedule rate for one measurement period is next listed using the Fig. 2 chart as a guide. For an I/O rate between R1 and R2, one rebuild is scheduled; upon measuring an I/O rate between R3 and R2, two rebuilds are scheduled; a measured I/O rate between R4 and R3 results in three rebuilds being scheduled; a measured I/O rate between R5 and R4 results in four rebuilds being scheduled while lower I/O rates than R5 result in five rebuilds being scheduled. The maximum number of scheduled rebuilds is five; any number can be used as the maximum. In the illustrated embodiment, a minimum size rebuild is one track. The information represented by the Fig. 2 chart is maintained for the parity array in the computer means for effecting scheduled rebuilds.

Fig. 3 illustrates reading data from one of the parity arrays 13-15, detecting rebuild criteria and priming the Fig. 1 illustrated system for rebuilding data. A usual read operation occurs at machine step 35 as initiated by other machine operations. At machine step 36 controller 11 (disk drives may contain error detecting facilities as well as the controller or host processors) detects errors in the data read from any of the disk drives 20-23; such errors are attempted to be corrected in a usual manner. At machine decision step 37, controller 11 determines whether or not the error corrections by the error redundancies in the individual

disk drives 20-23 were successful and whether or not fault tolerance was degraded even with a successful error correction. If error corrections were successful (high quality redundancy may still be indicated for some purposes and degraded redundancy may be indicated for other purposes, as will become apparent), then, assuming fault tolerant redundancy is not degraded for requiring a rebuild (NO degradation detected in machine step 37), machine operations proceed to other operations; no rebuild activity is indicated. On the other hand, if any one of the disk drives did not yield correctable data errors, which include a failure to respond, fault tolerance degradation is indicated. With the parity disk P 24, such data errors can still be corrected by reading the parity redundancy of the block from disk drive P 24, then computing the correct data from the data successfully read from the other drives and the parity redundancy. To achieve this known parity correction, the parity block stored in disk drive P 24 is read into controller 11. Then the data are corrected in machine step 39 using the known parity correction procedures. Such correction can occur in either a host processor 10 or in controller 11. At this point in time, the redundancy for the data unit being read has been removed. Then, at machine step 40, the parity correcting unit (host processor 10/controller 11) determines whether or not the parity correction is successful. Whenever the parity correction is unsuccessful, a subsystem error is flagged in a usual manner. Then, recovery procedures beyond the present description are required. Whenever the parity correction is successful, then at machine step 41, if it is determined that there is insufficient degradation of the fault tolerance effecting redundancy, other machine operations are performed; if it is determined that fault tolerance is unacceptable (a disk has failed, for example), then a rebuild is indicated.

Figure 4 illustrates the concepts of maintaining the desired redundancy without undue interference with day-to-day operations. The general arrangement of Fig. 4 can be thought of as establishing an interrupt driven examination of rebuild needs in a system. Machine step 45 represents monitoring an indicating I/O (input output) rate of operations for each parity group 13-15 of disk drives. At predetermined times, as will become apparent from Fig. 9, from such rate monitoring and indicating, a rebuild need is detected at machine step 46. Such detection may merely be a global rebuild flag or any entry in any of the Fig. 8 illustrated bit maps. If a rebuild is needed, then at machine step 47 a later described variable rate rebuild is scheduled. If a rebuild is not needed, then other machine other operations are performed.

Similarly, machine step 50 represents monitoring for idle time for idle time in any one of the parity groups 13-15. If idle time is detected, such as no pending access requests nor free standing operations are being performed, then machine step 51 represents detect-

ing a rebuild need. When a rebuild need is detected, then at machine step 52 a later-described idle time rebuild is effected. If no rebuild is required, other machine operations ensue.

Likewise, machine step 55 represents monitoring for a failed access, read operation or write operation in any one of the parity groups 13-15 or any access to a known failed drive. Upon detecting such an error, a rebuild need may be indicated as described for Fig. 3. Then at machine step 56 the rebuild needs are detected. On one hand, if the parity correction described in Fig. 3 was successful, a rebuild may be deferred, then from machine step 56 other operations ensue. If a rebuild is required, then the later-described opportunistic rebuild operations of machine step 57 are performed.

It is to be appreciated that the Fig. 4 illustration is tutorial; actual practical embodiments may differ in substantial details. Interleaving of the rebuild strategies can follow several variations. The determination of when and the total extent of a rebuild may substantially affect a given design.

Fig. 5 shows one way of selecting between two of the three illustrated data rebuild techniques. The selection procedure is entered at path 60 from other machine operations based upon any one of a plurality of criteria, such as a time out, time of day, number of accesses, the later-described rebuild schedule of the variable rate rebuild, whether or not bit map of Fig. 8 indicates any rebuild need and the like. Such a selection could typically reside in a dispatcher or other supervisory program (not shown). At machine decision or branching step 61, the type of rebuild needed to be evaluated is selected. Machine step 61 represents a program loop function controlled by software counter 62. Entry of the procedure at path 60 resets counter 62 to a reference state, counter 62 enables the decision step 61 to first evaluate an idle rebuild at machine step 65 as detailed in Fig. 10. If none of the parity arrays 13-15 are idle or there is no need for any rebuild (bit maps of Fig. 8 are all zeros), then operations return to machine step 61, counter 62 is incremented to a next value. This next value causes decision step 61 to effect evaluation of a variable rate rebuild at machine step 66 as detailed later in Fig. 9. The rebuild scanning may return to Fig. 5 from Fig. 9 to re-execute machine step 61 and increment counter 62. Other rebuild procedures may be employed (not described) as represented by numeral 67. Again, upon completing the rebuild evaluation, machine operations returning to the Fig. 5 procedure results in another incrementation of counter 62 and execution of machine step 61. Since the program loop scanning of the procedures has been completed, other machine operations are performed as indicated by numeral 68. The order of scanning the rebuild procedures or methods is arbitrary. As shown in Fig. 11, the opportunistic rebuild procedure is always entered from a disk accessing

operation. Any method of scanning rebuild procedures may be employed for selecting any one of a plurality of rebuild procedures.

Fig. 6 is a diagrammatic plan view of a disk in any of the disk devices 20-24; a plurality of such disks are usually stacked to be coaxially co-rotating in the respective devices. All tracks on the respective disks having the same radial location constitute a cylinder of such tracks. When employing a traditional fixed block architecture, each disk 70 may be envisioned as a plurality of disk sector-indicating radially-extending machine-sensible lines 71. Each disk sector between the lines is addressable as a data-storing unit. In a count-key-data (CKD) disk a single radially-extending track index line is used. Each disk has a multiplicity of addressable circular tracks, or circumvolutions of a spiral track, residing on each disk 70. A track 72 may be affected by error requiring a partial rebuild of the array. In disk 70 the data contents of the error affected track 72 may be reassigned to track 73; in a rebuild the data contents of all tracks 72 in the respective disk device 20-24 are similarly reassigned to their respective tracks 73. In one mode, the data contents of a cylinder of tracks in which track 72 may be reassigned to a cylinder of tracks including track 73. In another mode, only the contents of a single track are reassigned. When a disk device is totally replaced, then the data from all of the remaining devices 20-24 are used to compute the data for the replaced disk. The decision when to replace a disk device that is partially operable may be based upon the number of unusable or bad tracks on the device, the character of the error causing failure, and the like.

Fig. 7 illustrates a data structure for one implementation of the variable rate rebuild method. The objective of this method is to maintain during the rebuild period at least a minimum level of subsystem performance, i.e. response time to received I/O requests. Three registers or data storage locations 80-82, either in a host processor 10 or controller 11, store control information need to effect the variable rate rebuild method in the respective parity arrays 13-15. Each register 80-82 is identically constructed, the criteria information may be different to accommodate arrays having different performance characteristics or system usages. Field rate 83 stores a number indicating the rate of rebuild, i.e. one rebuild per second, two per second, etc. Field AVG-IO 84 stores the average I/O response time, possibly expressed in terms of its corresponding I/O request rate, in a predetermined measuring period. The I/O response time or request rate is used to compute the rebuild rate. Fields 85-88 respectively store rebuild rates for the I/O request rates T-1 through T-4 for various rebuild rates, no limitation thereto intended. The total number of disk accesses to a parity array is an indication of response; the greater the number of access requests, the lower the desired rebuild rate. Thresholds T-1 through T-4

correspond to decreasing numbers of access requests rates and indicate higher and higher rebuild rates. Threshold T-1 indicates an access rate greater than which would result in no rebuild being permitted by the variable rate rebuild method. Threshold T-2 indicates an access rate greater than T-3 and smaller than T-1 and which permits one rebuild access (i.e. rebuilds one track of data) during a constant request rate measuring period. Similarly, threshold T-3 indicates an access rate greater than T-4 and smaller than T-2 and which permits two rebuild accesses during the constant rebuild rate measuring period. As request rates continue to decrease, corresponding increases in rebuild rates occur. A predetermined maximum rebuild rate for the system may be established. In another implementation of the variable rate rebuild method, an average response time can be directly measured during each successive measuring period. If the measured response time is slower than a desired response time, the rebuild rate to be used during the next successive measuring period is reduced. If the measured response time is shorter than the desired response time, the rebuild rate used in the next successive measuring period is increased. Alternatively, if I/O access queues exist, then the rebuild rate may be selected to be inversely proportional to the length of the access queues for the respective parity arrays. Any of the above-described measurement techniques may be employed for establishing the rebuild rate control information stored in fields 83 and 84. Numeral 89 indicates that additional criteria may be used for determining a rebuild rate which is inversely proportional to accessing/free-standing array operations.

Information specifying which tracks need rebuilding is maintained in bit maps 95-97 (Fig. 8) respectively for parity arrays 13-15. The rows 105, 106, 107 ... of bit-containing squares 99 respectively indicate sets of logical tracks on one recording surface of disk 70 (Fig. 6). The columns 100, 101, 102 ... of bit-containing squares 99 respectively represent logical cylinders of the logical tracks. Each logical track includes one physical track in each device 20-24 and each logical cylinder includes one physical cylinder in each device 20-24. When any one of the parity groups 13-15 is providing complete redundancy, then all of the squares 99 in the respective bit map 95-97 contain binary 0's. Any track needing a rebuild, whether as part of a complete rebuild of a disk device or a partial rebuild, is indicated by a binary 1 in the square or bit position 99 of the respective bit map. Scanning the bit maps for ascertaining rebuild needs follows known techniques. An index or written-to-bit-map value (not shown) may be used to indicate that a respective bit map either contains at least a binary 1, all binary 0's or the number of binary 1's in each respective bit map.

Fig. 9 illustrates one implementation of the variable rate rebuild method, including generating the control information for the Fig. 7 data structure. The

description is for one of the parity arrays; modifying the machine operations to accommodate a plurality of parity arrays can be achieved in any one of several approaches. For example, only one of the three arrays 13-15 may have a non-zero bit map; then only the array indicated by the non-zero bit map is processed. If a plurality of bit maps are non-zero, priority of rebuilding in the three arrays can be based upon the least busy array, a designation of relative importance of the arrays to continued successful system operations, and the like. In any event, entry of the Fig. 9 illustrated machine operations is over path 110 from either Fig. 4 or 5, the present description assumes entry from Fig. 5. At machine step 111 whether or not a measurement period has timed out is sensed. If a measurement period did not time out, then at machine step 112 the access tally is updated (other rebuild rate indicating criteria may be updated as well). Following the update, at machine decision step 113 field 83 is examined along with an elapsed time indication. Elapsed time is computed from the time of day the last rebuild for the parity error was completed, such as indicated by numeral 89 of Fig. 7. By way of design, such last rebuild is the last rebuild by any of the methods being used, the last rebuild of either the idle or variable rate rebuild method or the last rebuild achieved by the variable rate rebuild method. If the time for rebuild has not been reached, then the Fig. 5 illustrated scanning procedure is reentered. If a rebuild is to be scheduled, then at machine step 118 the cylinder and track(s) to be rebuilt are selected. In making the cylinder and track selection, it is desired to minimize seek times to reach a track(s) to be rebuilt. For a striped mode array, since the track access mechanisms (not separately shown) of the drives in each parity array always have a common radial position over disks 70 of the respective devices 20-24 (scanning a track in a current logical cylinder comprising all physical cylinders in the parity group devices at a same radial position), a track in the current cylinder or in a cylinder having a closest radial proximity to the current cylinder is selected. Such track(s) is identified by analysis of the array Fig. 8 bit map. This analysis is straight forward and is not further described.

For a parity array operating in the independent mode, the same general approach is used. Each of the devices in the independent mode may be scanning tracks/cylinders having different radial positions. The track/cylinder to be rebuilt is the cylinder that is located at a mean radial position between a first device in the array having its access mechanism at a radial inwardmost position and a second device having its access mechanism at a radially outwardmost position of all the devices in the array. For example, if the first device has its access mechanism for scanning track 72 of Fig. 6, the second device has its access mechanism for scanning track 73 and the other two operational devices in the parity array have their re-

spective access mechanisms positioned radially for scanning tracks radially intermediate between tracks 72 and 73, then the cylinder radially midway between tracks 72 and 73 is examined for rebuild. If the midway cylinder has no track to be rebuilt, then adjacent cylinders are successively examined at successively increasing radial distances from the midway cylinder. This determination follows examining the cylinder which is next radially outward of the midway cylinder, thence if that cylinder has no track requiring rebuild, then the next radially inward cylinder is examined, etc.

After selecting the cylinder in machine step 118, a track(s) in the selected cylinder is rebuilt in machine step 119. This rebuilding involves computing the data from the corresponding physical tracks in the other devices of the array; then storing the computed data into the selected track(s). Upon completion of the rebuilding, the respective bit position(s) of the array bit map of Fig. 8 is reset to 0. Then at machine step 120 a check is made on whether or not the parity array is idle and there are still tracks to be rebuilt. Whenever the array is idle and a track of the parity array still needs to be rebuilt, machine steps 118 and 119 are repeated until the array is no longer idle or all rebuilds have been completed, then other machine operations are performed.

Calculation of the variable rebuild rate occurs whenever the machine step 111 indicates the measurement period has expired. In one embodiment of computing the desired rebuilt rate, at machine step 125 the number of accesses to the parity array are averaged to obtain an average access rate. Such averaging allows varying the measurement period without affecting other variables. The average rate is stored in field 84. Then the access tally is reset in machine step 126, such access tally may be stored in a register 80-82 as represented by numeral 89. In machine step 127 the rebuild rate is determined by comparing the access tally values in fields 85-88 with the field 84 value. Then the rebuild rate corresponding to the threshold field 85-88 having a value least less than the field 84 value is stored in field 83 as the new rebuild rate. Remember that T-1's rebuild rate is zero. If a queue length criterion is used, then the queue length(s) are examined and the rebuild rate corresponding to the respective queue lengths is selected. Of course, algorithmic calculations can be used rather than table lookup; such calculation results are rounded to a next lowest time slice or unit value.

Fig. 10 illustrates one embodiment of the idle rebuild method. Entry into the method is from Fig. 4 or over path 129 from Fig. 5. Whether or not the parity array is idle and a track in the parity array needs a rebuild is checked at machine step 130. If the parity array is not currently idle or there is no need for a track rebuild, then the operation returns over path 135 to the caller, such as the Fig. 5 illustrated selection method. When the parity array is idle with a rebuild need,

then at machine step 131 a cylinder and one of its tracks are selected for rebuild. This selection uses the
aforedescribed selection method. Following selection
and location of the selected track, machine step 132
rebuilds the track contents. Machine step 133 then
checks to see if the parity array is still idle, if yes steps
131 and 132 are repeated until either no more rebuilds
are needed or the parity array becomes busy. At that
point other machine operations ensue.

Fig. 11 illustrates the opportunistic rebuild method. A track access operation is initiated over path 139.
For purposes of illustration, the ensuing machine step
140 is an attempted read from the track to be accessed.
Any access operation may be used, such as write, format
write, verify, erase and the like. Assuming a read operation,
machine step 141 determines whether or not a hard (uncorrectable)
error has occurred. Included in the machine step 141 operation
is detection that the device containing the accessed track
is already known to have failed and a rebuild is pending
or already in progress for that device. If the read produced
no hard error, i.e. no errors detected or a corrected error
occurred, machine step 142 checks the quality of the read-back
operation. Since a corrected error may not be repeated,
machine step 142 may not invoke the opportunistic rebuild
method, choosing to proceed as OK over path 143 to other
operations. If the quality requirements are not met, such
as determinable by evaluating read signal quality, the type
and extent of the corrected error, systems requirements
(desired quality of the redundancy in the array) and the like,
the opportunistic rebuild is initiated (the NO exit from step 142).
Machine step 144 effects a rebuild of the currently accessed
track from either machine step 141 detecting a hard error or
from machine step 142. Rebuilding follows the above-described
method of rebuild. Upon completing rebuilding the accessed
track data in machine step 144, in machine step 145 a check
is made on whether or not the parity array is idle. If the
parity array is not idle, then other operations are performed;
if the parity array is idle, then machine step 146 rebuilds a
next track. Such a rebuild includes selecting a cylinder and
track followed by the actual rebuild method. Machine steps
145 and 146 repeat until either no more rebuilds are needed
(bit maps are all zeros) or the parity array becomes active.

Claims

1. A method of automatically maintaining fault tolerance in a parity array (13) consisting of a plurality of disk drives (20, 21, 22, 23, 24), comprising the machine-executed steps of:
 - detecting a degradation in the fault tolerance of the parity array;
 - evaluating the current information handling activity of the parity array;

establishing a plurality of data rebuild methods for removing the fault tolerance degradation from the parity array; and

analyzing the current information handling activity of the parity array and selecting one of the plurality of rebuild methods to ensure that the performance of said current information handling activity is not degraded below a predetermined level.

2. A method as claimed in claim 1, wherein one of said plurality of data rebuild methods comprises the machine-executed steps of:

- measuring the rate of machine operations of the array;

- establishing the rate of rebuilding data affected by the error in inverse proportion to the measured rate of machine operations; and

- rebuilding the data accordingly.

3. A method as claimed in claim 1 or 2, wherein one of said plurality of data rebuild methods comprises the machine-executed steps of:

- detecting that the parity array is idle;

- detecting that one of the disk drives is affected by error;

- rebuilding the data during said detected idle times.

4. A method as claimed in claim 3 as dependent on claim 2, further comprising the machine-executed steps of:

- completing a data rebuild using the rebuild method of claim 2;

- detecting that the parity array is idle; and

- continuing the data rebuilding of additional addressable error-affected data so long as the parity array is idle.

5. A method as claimed in any of claims 1 to 4, where one of said plurality of data rebuild methods comprises the machine-executed steps of:

- performing a data area access operation to an addressable data unit in the parity array;

- while performing the data area access operation, detecting that the addressed data unit is affected by error and in need of a data rebuild;

- rebuilding the addressable data unit being accessed.

6. Apparatus for automatically maintaining fault tolerance in a parity array (13) consisting of a plurality of disk drives (20, 21, 22, 23, 24), the apparatus comprising:

- means for detecting a degradation in fault tolerance in the parity array such that one of the disk drives needs to have data rebuilt;

means for measuring a rate of machine operations of said array;

means responsive to said measured operations rate for establishing a rate of rebuilding for the array for recovering from said degradation; and

a plurality of data rebuild means for performing data rebuild in said one disk drive.

7. Apparatus as claimed in claim 6, further comprising:

control means (11) for controlling access to the disk drives in the parity array and for detecting when the array is currently not being accessed for a data handling operation.

8. Apparatus as claimed in claim 7, wherein one of said plurality of data rebuild means is connected to said control means and to said rate establishing means for determining whether a rebuild can be scheduled, and if so, activating the control means to give access to the parity array to perform a series of rebuild operations at the established rate.

9. Apparatus as claimed in claim 8, wherein said rate establishing means further comprises:

means for indicating a plurality of rebuild rates, each corresponding to predetermined ranges of machine operations rates and in inverse proportion to the operations rate; and

means for indicating one of said plurality of rebuild rates as said established rate corresponding to the current machine operations rate.

10. Apparatus as claimed in claim 7, 8 or 9, wherein one of said plurality of data rebuild means is connected to said control means and responsive to the parity array not being currently accessed to activate the control means to give access to the parity array to perform a data rebuild.

11. Apparatus as claimed in any of claims 7 to 10, wherein one of said plurality of data rebuild means is connected to the control means and responsive to a disk drive access occurring in an area of the array that needs a data rebuild to perform a data rebuild.

12. Apparatus as claimed in any of claims 7 to 11, wherein each of the data rebuild means upon completing a data rebuild operation is responsive to the control means indicating no current access or no pending access requests to initiate another data rebuild operation.

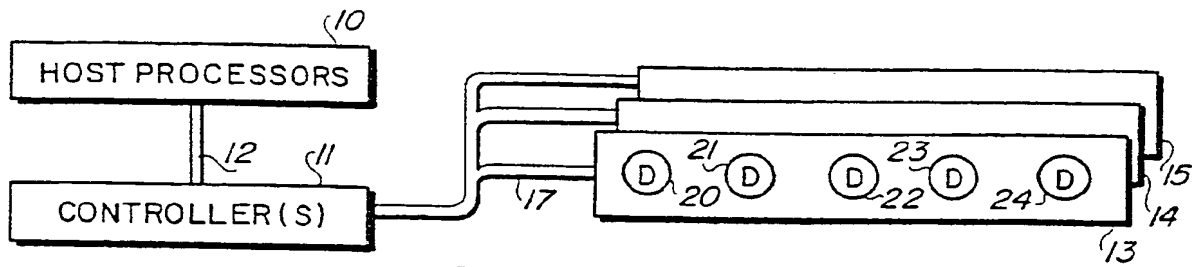


FIG. 1

AVERAGE
RESPONSE
TIME

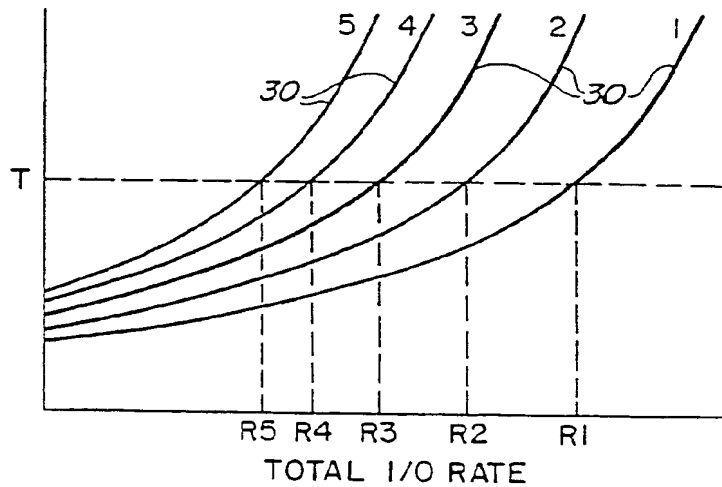


FIG. 2

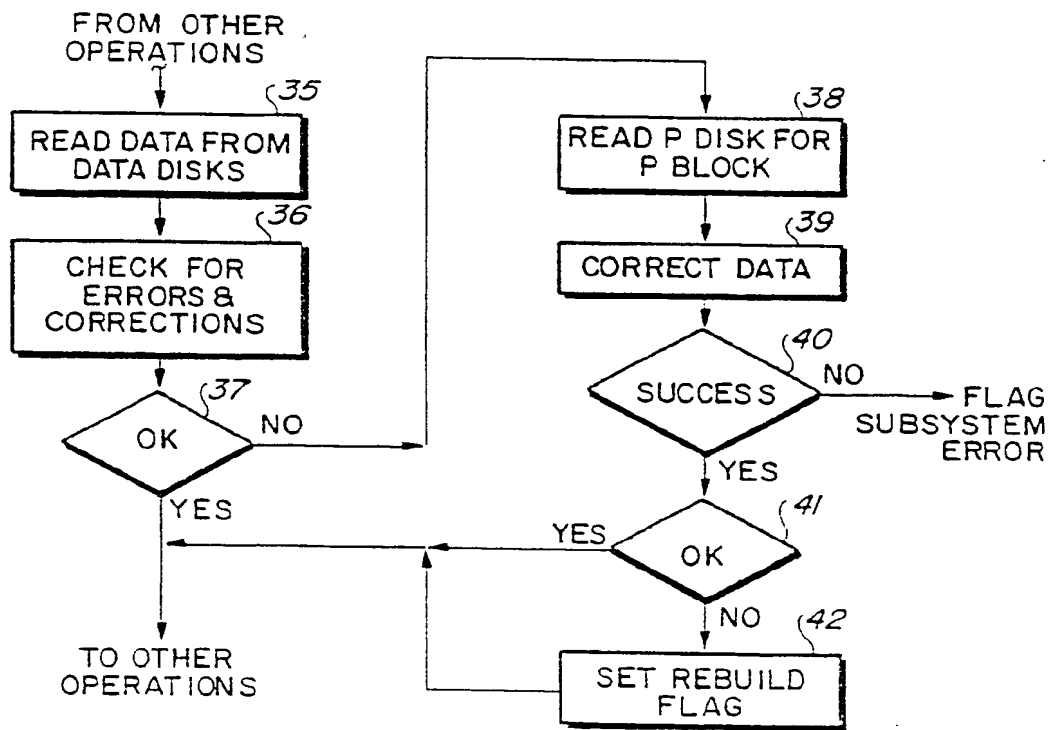


FIG. 3

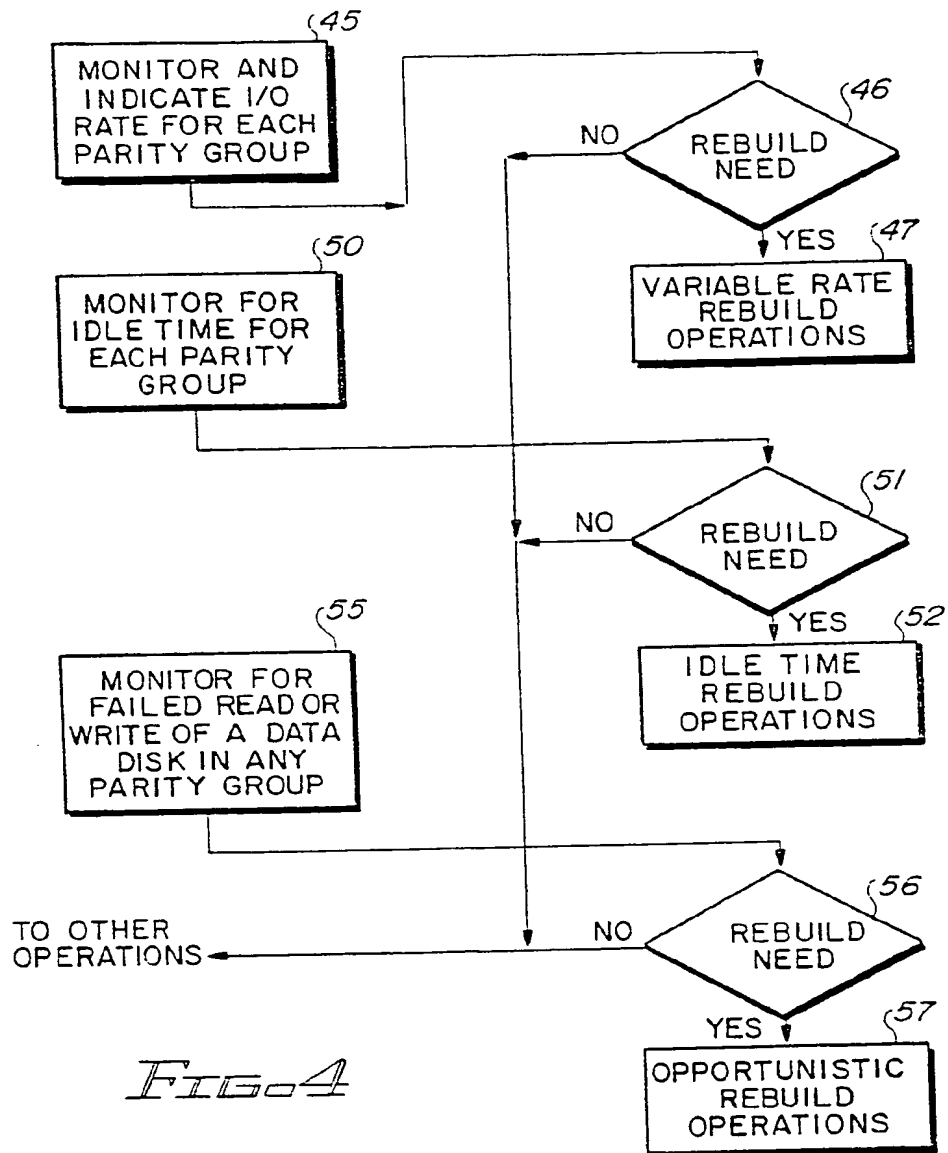


FIG. 4

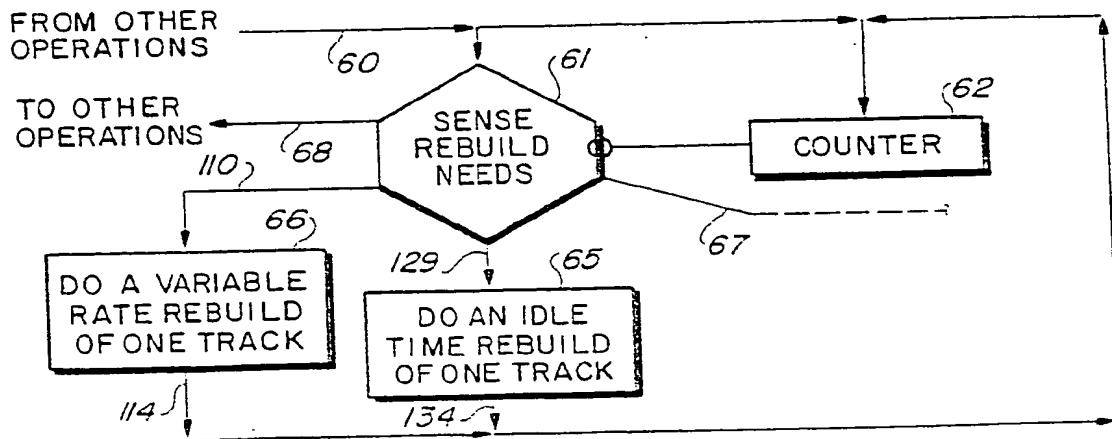


FIG. 5

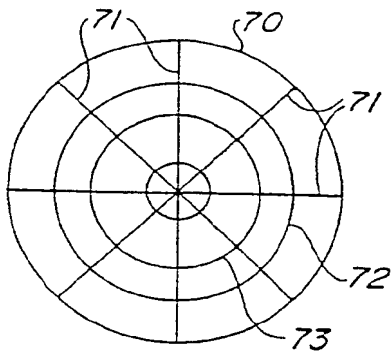


FIG. 6

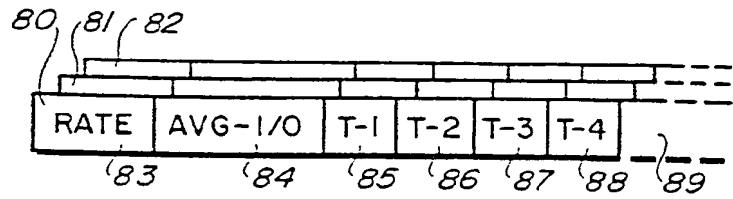


FIG. 7

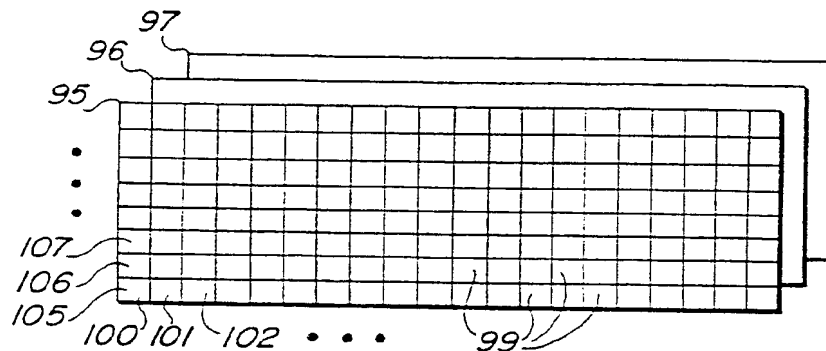


FIG. 8

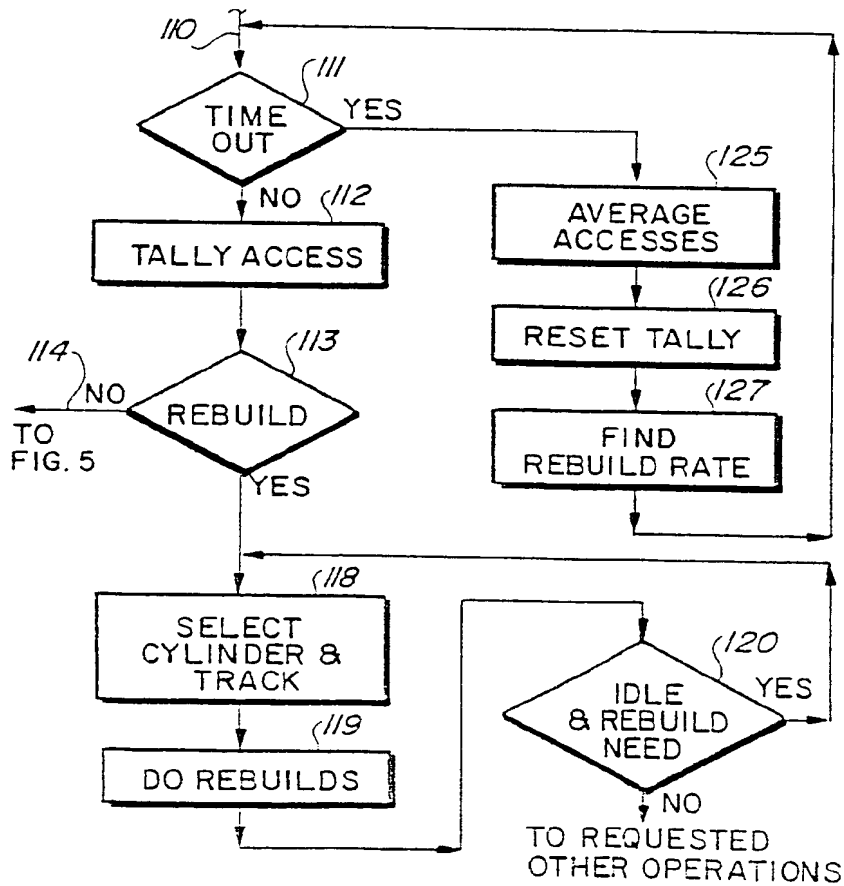


FIG. 9

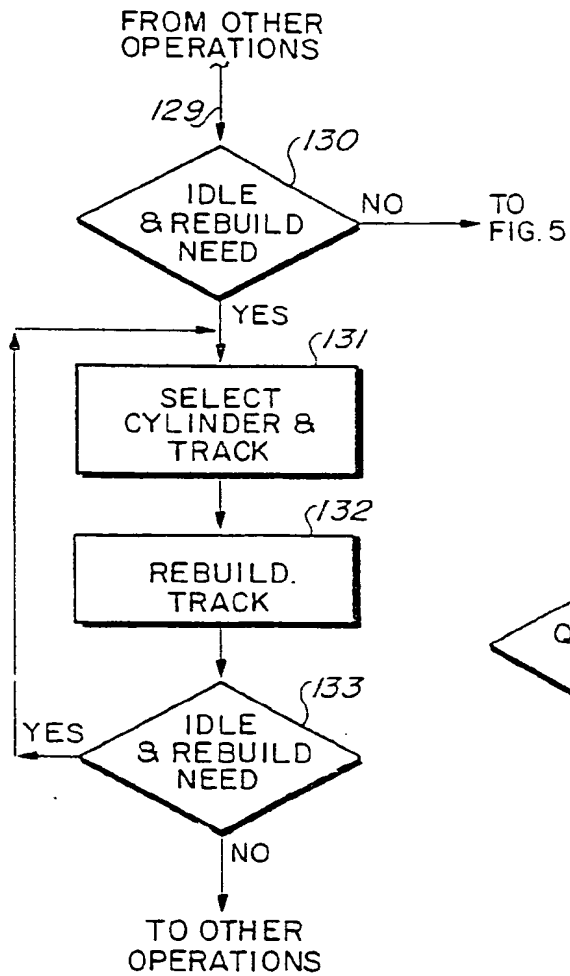


FIG. 10

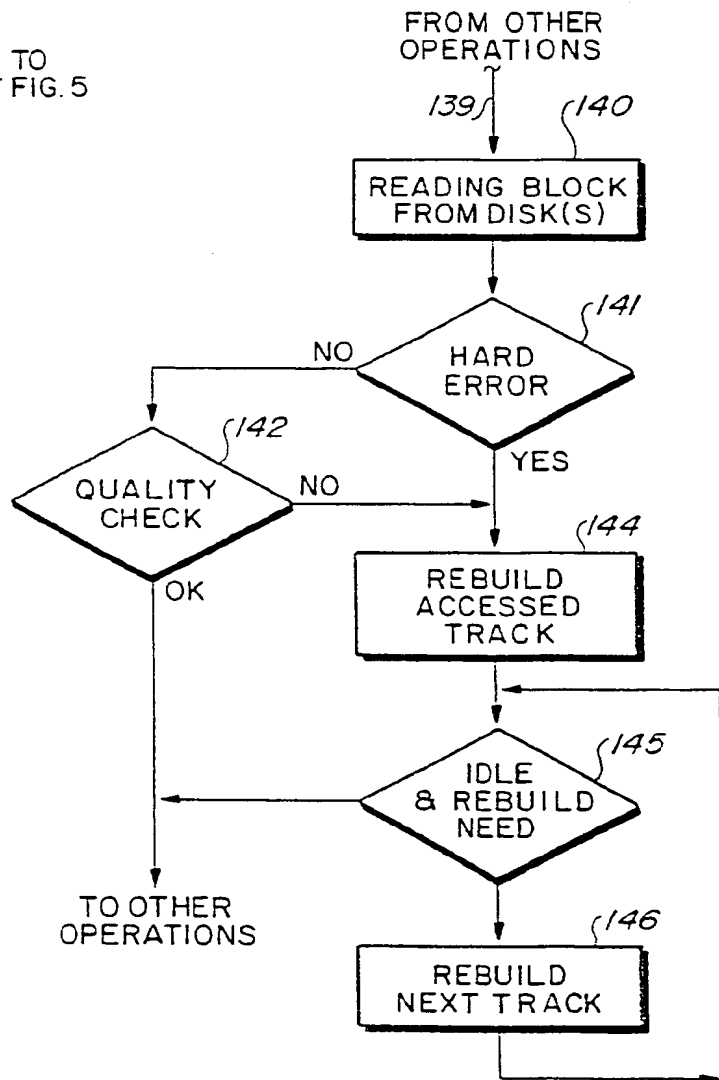


FIG. 11



(11) Publication number : **0 519 670 A3**

(12)

EUROPEAN PATENT APPLICATION

(21) Application number : **92305481.1**

(51) Int. Cl.⁵ : **G11B 20/18, G06F 11/10**

(22) Date of filing : **15.06.92**

(30) Priority : **18.06.91 US 717263**

(43) Date of publication of application :
23.12.92 Bulletin 92/52

(84) Designated Contracting States :
DE FR GB IT

(88) Date of deferred publication of search report :
03.03.93 Bulletin 93/09

(71) Applicant : **International Business Machines Corporation**
Old Orchard Road
Armonk, N.Y. 10504 (US)

(72) Inventor : **Ng, Spencer Wah-Fung**
974 Hampswood Way
San Jose, California 95120 (US)
Inventor : **Palmer, David William**
248 Omira Drive
San Jose, California 95123 (US)
Inventor : **Thompson, Richard Scott**
67 Dearwell Way
San Jose, California 95138 (US)

(74) Representative : **Blakemore, Frederick Norman**
IBM United Kingdom Limited Intellectual Property Department
Hursley Park
Winchester Hampshire SO21 2JN (GB)

(54) **Apparatus for maintaining fault tolerance in a disk drive array.**

(57) Fault tolerance in a redundant array 13 of disk drives 20, 21, 22, 23, 24 is degraded when errors exist in the array. Data in the array is rebuilt to remove the degradation, either for entire disk drives or for partial data, with minimal impact on array resources. Rebuilding occurs during idle time of the array or is interleaved between current data area accessing operations of the array at a rate which is inversely proportional to the activity level of the array. Alternatively, data is rebuilt when a data area being accessed is in need of rebuilding.

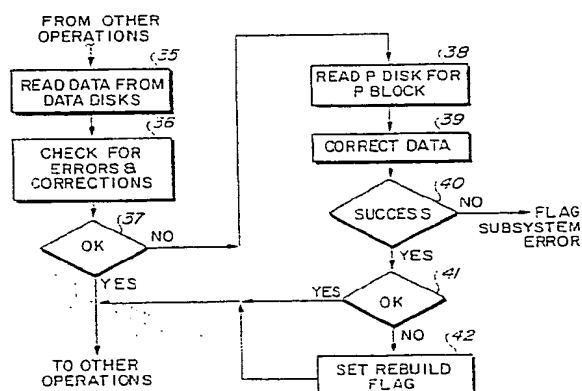


FIG. 5

EP 0 519 670 A3



European Patent
Office

EUROPEAN SEARCH REPORT

Application Number

EP 92 30 5481

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl.5)
A	EP-A-0 320 107 (MICROPOLIS CORPORATION) * column 17, line 19 - column 22, line 17; figure 7 *	1,6	G11B20/18 G06F11/10
A	EP-A-0 274 817 (SEAGATE TECHNOLOGY INTERNATIONAL)		
A	EP-A-0 426 185 (COMPAQ COMPUTER CORPORATION)		
A	WO-A-8 809 968 (CAB-TEK INC)		
			TECHNICAL FIELDS SEARCHED (Int. Cl.5)
			G11B G06F
The present search report has been drawn up for all claims			
Place of search THE HAGUE		Date of completion of the search 23 NOVEMBER 1992	Examiner DEVERGRANNE C.
<p>CATEGORY OF CITED DOCUMENTS</p> <p>X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document</p> <p>T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons ⌘ : member of the same patent family, corresponding document</p>			

EP0 FORM 1501 01/82 (P0001)